and that the majority (90-99%) of these have little in common with known biologically extant compositions.

The aim of the present invention is to attempt to alleviate some of the above described problems and to reduce the large number of irrelevant compositions returned by existing tools.

Any discussion of documents, acts, materials, devices, articles or the like which has been included in the present specification is solely for the purpose of providing a context for the present invention. It is not to be taken as an admission that any or all of these matters form part of the prior art base or were common general knowledge in the field relevant to the present invention as it existed before the priority date of each claim of this application.

Summary of the Invention

In a first broad aspect, the present invention incorporates statistical measures of biological relevance for the candidate compositions returned.

Typically, biological relevance, expressed as a numerical score, or biological index, is determined by statistical comparison to an established reference set of known and fully characterised compositions, in the case of glycans a reference set such as the Glycosuite (http://www.glycosuite.com) database. The biological index of any given composition may then be used as a basis for discarding biologically "unlikely" compositions, as well as for ranking (sorting) of returned compositions by biological likeliness.

Empirically, for glycans this allows between 90-99.9% of candidate compositions returned by any given search to be discarded, whilst preserving and ranking the remaining, biologically likely compositions.

In one aspect the present invention provides a method of determining the likelihood of a saccharide composition of a candidate glycan comprising:

providing a search mass of a glycan whose composition is to be determined;

generating a list of possible glycans made up of components, including monosaccharides, whose total mass is within a predetermined tolerance of the search mass;

selecting a reference group of known characterised glycans ;

establishing the mean and standard deviation of each component appearing in the reference group of the known characterised glycans ;

for each candidate glycan calculating a partial score for each component in that theoretical glycan candidate, the partial score being calculated from the mean and standard deviation of the component appearing in the reference group and which provides a measure of the likelihood of that component being present in the candidate glycan; and

combining the partial scores to provide an indication of the likelihood of that candidate glycan occurring.

More particularly, for glycans, in one aspect the present invention provides a method of characterising glycans comprising the steps of:

necessary in order to obtain a sufficiently large sample size (preferably at least 100 known compositions). In the case of the Glycosuite database of known sugar structures, a mass tolerance of 200 Da was empirically determined to be sufficient to provide in excess of 100 known compositions for search masses up to around 3500.

By way of example if the search mass were 1000 Da there may be 100 known glycans in the database whose mass is between 800 and 1200 Da. The mean and standard deviation of each of every monosaccharide/component appearing in those known glycans in the database is then determined. If we take HexNAc as an example we may find that, on average, the 100 known glycans contain 3.3 HexNAc monosaccharides with a standard deviation of 2.3. This process is repeated to calculate the mean and standard deviation for each monosaccharide component Hex, dHex, pent et al, and each adduct in the known glycans, if adducts are being accounted for.

For each candidate glycan composition "Partial scores" are then determined from the means and standard deviations calculated above. These are calculated for each monosaccharide in the given composition as the absolute value of the difference between the mean number of that monosaccharide in the reference set and the observed number of that monosaccharide in the theoretical candidate composition, divided by the standard deviation of that monosaccharide in compositions from the reference set. ie:

$$partialscore_{monosac} = \frac{|mean_{monosac} - observed_{monosac}|}{stdev_{monosac}}$$

where $mean_{monosac}$ is the mean number of the given monosaccharide in the reference data set (Glycosuite); $observed_{monosac}$ is the number of the given monosaccharide in the theoretical candidate composition; and $stddev_{monosac}$ is the standard deviation of the given monosaccharide in the reference data set.

By way of example if the theoretical glycan composition includes two HexNAc, three Hex and 1 NeuAc, the partial score for each of those three monosaccharides is calculated for that theoretical candidate glycan composition. Partial scores need not be calculated for monosaccharides which do not appear in the candidate theoretical glycan composition.

In the event that the $mean_{monosac}$ equals the $observed_{monosac}$ for a particular glycan, the system is arranged to give the partial score a minimum value of 0.01.

Thus, the partial score of a monosaccharide is in fact the number of standard deviations the number of away from the mean that that monosaccharide is in the